No Machine Can Universally Pass the Turing Test

Humor as the Final Test of Person-Specific Alignment

A conceptual account of high-level subset alignment
under the strict single-judge condition

Jia, Baolong

Independent Researcher

Bilingual Conceptual Paper ·Draft v0.4 ·July 2026

Contents

Abstract . . . 1
1. The Wrong Question: Can a Machine Generate a Joke? . . . 1
2. Three Judgments That Must Not Be Confused . . . 2
3. Functional Approximation and Subset Alignment . . . 2
4. What "Complete Alignment" Means . . . 3
5. Humor as the Momentary Intersection of Two Internal Spaces . . . 3
6. "Spark" and the Failure of Template Humor . . . 4
7. Why N=1 Changes the Turing Test . . . 4
8. Why Humor Is a Decisive N=1 Probe, Not the Verdict . . . 5
9. Conclusion . . . 6
References . . . 6
摘要 . . . 1
1. 错误的问题：机器能生成笑话吗? . . . 1
2. 三种绝不能混淆的判断 . . . 1
3. 功能近似与子集对齐 . . . 2
4. "完全对齐"究竟是什么意思 . . . 2
5. 两个内部空间的瞬间交集 . . . 3
6. "闪光"与模板幽默的失败 . . . 3
7. 为什么 N=1 改变了图灵测试 . . . 3
8. 为什么幽默是重要的 N=1 探针，而不是判决本身 . . . 4
9. 结论 . . . 4
参考文献 . . . 5

Abstract

Artificial intelligence is usually evaluated by function: whether it can answer, predict, plan, imitate, or generate. Humor is different. A humorous expression does not succeed merely because it has the recognizable structure of a joke. It succeeds only when a speaker and a particular listener become aligned, at one moment, over a narrow but high-level subset of their internal conceptual spaces.

This paper calls that condition **high-level subset-parameter alignment**. The relevant subset may contain cultural memory, personal experience, unspoken expectations, linguistic instinct, social distance, emotional state, taboo boundaries, conversational rhythm, and mutual knowledge of the present relationship. The two minds do not need to be identical as wholes. But within the subset that decides the humorous event, the alignment must be complete: one wrong implication, one unnatural word, or one misplaced beat can destroy the effect.

The paper distinguishes three judgments that must never be conflated. **Humorous success** means that an expression makes the judge laugh. **Local N=1 passage** means that, after sustained interaction, the sole judge cannot reliably distinguish the AI from a human. **Universal passage** means that the same system can achieve local passage before any possible sole judge. Laughter is neither necessary nor sufficient for Turing passage. A judge may laugh repeatedly while still recognizing the machine; humor has succeeded, but the Turing test has failed. A human may also tell an unfunny joke; humor has failed, but this alone does not identify a machine.

Humor matters because it can become a probe. If a sole judge repeatedly detects templated structure, unnatural phrasing, or failed person-specific alignment and can use those features to identify the AI reliably, the system fails that judge's N=1 test. The same system may nevertheless pass before another judge. It can therefore pass locally without passing universally. The claim is not that machines can never output funny sentences, nor that every failed joke proves a machine. It is that humor exposes judge-dependent alignment, preventing one local success from becoming a universal passing certificate.

Keywords: humor; N=1 Turing test; artificial intelligence; high-level alignment; subset parameters; human-AI interaction; comic intuition

---

1. The Wrong Question: Can a Machine Generate a Joke?

The easiest way to misunderstand machine humor is to ask whether a machine can generate a joke.

It can. A system trained on enough language can reproduce the visible forms of humor: setup and punchline, reversal, exaggeration, wordplay, incongruity, self-deprecation, and irony. It can also combine these forms into sentences that some people find funny. None of this reaches the hardest part of humor.

A joke is a linguistic object. Humor is an event between minds.

The object can be copied, stored, translated, and rated. The event depends on who speaks, who listens, what they share, what remains unspoken, what has just happened, what must not be said directly, and how long the speaker waits before saying it. The same sentence can be brilliant, ordinary, cruel, or meaningless under different relations.

Therefore, the real question is not:

> *Can the system construct something with the form of a joke?*

It is:

> *Can the system find the exact hidden region in which this particular listener will experience the expression as both unexpected and right?*

That is not a general generation problem. It is a person-specific alignment problem.

2. Three Judgments That Must Not Be Confused

The word "success" refers to three different outcomes in discussions of AI humor.

Humorous success occurs when the listener experiences amusement or laughter. It describes the effect of an expression.

Local N=1 passage occurs when one specified judge, after sustained and open-ended interaction, cannot reliably distinguish the AI from a human. It describes the result of one single-judge test.

Universal passage would require the same system to achieve local passage before every possible sole judge. It is a claim across different N=1 tests, not the result of one N=1 encounter.

The distinction can be made concrete.

- Judge A laughs repeatedly but knows that the speaker is AI. The AI has achieved humorous success but has failed Judge A's N=1 Turing test.
- Judge A laughs and, after sustained interaction, cannot reliably decide whether the speaker is human or AI. The AI has passed Judge A's N=1 test.
- Judge B detects mechanical or templated humor and uses it to identify the AI reliably. The same AI has failed Judge B's N=1 test.

The combined result is exact: the system passes for A and fails for B. It has not "passed the Turing test" without qualification, and it has not failed every possible test. It has passed one local test and failed another; therefore it has not passed universally.

Humor is a probe, not the verdict. A bad joke by itself proves nothing because humans also produce bad jokes. Humor becomes relevant to the Turing verdict only when its pattern gives the sole judge reliable grounds for distinguishing machine from human.

3. Functional Approximation and Subset Alignment

Many intelligent behaviors can be approached through functional approximation. The system is given a task, receives feedback, and improves its ability to produce the required result. There may be many acceptable internal routes. If the function is achieved, the approximation succeeds.

Humor has a different success condition. There is no listener-independent output called "the correct joke." The target changes with the recipient.

A system can approximate the general form of successful humor by learning what frequently makes people laugh. This produces population-level competence. It can learn that a setup creates one expectation and that a punchline replaces it with another. It can learn common comic topics, safe degrees of exaggeration, and familiar rhythms. This is why generated humor can be structurally complete and occasionally effective.

But a population pattern is an average across differences. Humor with one person depends on precisely those differences that the average removes.

The decisive issue is not whether the system has learned a comic function in general. It is whether the system and the listener are aligned on the small subset of parameters that matters now. That subset can

include:

- a memory shared only by the two participants;
- a private meaning attached to an ordinary word;
- an expectation that has never been stated;
- a personal boundary between boldness and offense;
- a rhythm natural to this person′ s speech;
- the emotional weight of a recent event;
- the listener′ s awareness that the speaker is trying to be funny;
- and the speaker′ s awareness that the listener is testing the attempt.

Functional approximation permits alternative solutions. Subset alignment does not. Within the active humorous subset, an apparently minor deviation can be decisive.

4. What "Complete Alignment" Means

The expression **complete alignment** can be misunderstood. It does not mean that speaker and listener must have identical minds. They may disagree about almost everything and still share a perfect joke.

Completeness applies only to the active subset.

Imagine that a humorous moment depends on six hidden conditions. Both participants must recognize the same earlier event; understand the same double meaning; treat the same norm as temporarily violable; perceive the same degree of intimacy; hear the same ironic tone; and expect the sentence to stop at the same point. If five conditions align but the sixth does not, the joke may fail completely.

This is why humor is often destroyed by something extremely small:

- a word is semantically defensible but no human would naturally use it there;
- a reversal is understandable but too visibly manufactured;
- an implication is almost right but attributes the wrong motive to the character;
- the line arrives after the listener has already predicted it;
- the speaker explains the relation that the listener needed to discover;
- the violation is recognized, but not as benign;
- or the listener feels the system aiming at a humor template rather than responding to the moment.

The failure is not proportional to the size of the error. One word can collapse the event because that word lies inside the decisive subset.

This is what makes humor a high-level alignment problem. Grammar and dictionary meaning can both be correct while the expression remains wrong. The relevant judgment occurs at the level of intention, relationship, naturalness, expectation, and shared implication.

5. Humor as the Momentary Intersection of Two Internal Spaces

"Vector space" is useful here as a conceptual image, not as a mathematical proof. A person′s mind contains a vast network of associations, memories, values, emotional weights, linguistic habits, and predictions. Another person has a different network.

Ordinary communication requires partial overlap. Humor requires something narrower and more exact: the two networks must intersect on the right high-level relations at the same moment.

The speaker first leads the listener toward an expected interpretation. The humorous turn then opens a

second interpretation. For the turn to work, the listener must be able to move from the first to the second without being told the whole route. If the second interpretation is too obvious, there is no spark. If it is too distant, there is only confusion. If it crosses the wrong boundary, there is offense. If it requires explanation, the timing is already lost.

The successful moment is therefore an intersection with tension. Speaker and listener must be aligned enough to share the hidden route, but the listener must not have completed the route before the turn arrives.

This is why the finest humor often feels impossible to paraphrase. The information content may be simple, but the effect lies in the path of recognition. An explanation gives the destination without reproducing the movement. It can show why the joke was constructed, but not recreate the instant in which two internal spaces met.

6. "Spark" and the Failure of Template Humor

AI humor often has structure without spark.

"Spark" does not mean randomness, complexity, or rare vocabulary. It is the instant in which an expression is unexpected before it appears and exact after it appears. The listener did not predict it, yet immediately feels that it belongs to this situation.

Template humor reverses that relation. Its structure is visible before its meaning becomes alive. The listener recognizes the machinery: a normal setup, an exaggerated contrast, and a final sentence engineered to oppose the opening. The result may still be interpretable and mildly amusing, but the mechanism arrives before the insight.

Large language models are especially capable of producing this kind of structurally correct humor because they learn recurring relations among linguistic forms. The same ability creates a recognizable weakness. When many candidate patterns are compressed into a generally safe response, the expression can sound like the statistical center of previous jokes. It fits the category but not the person.

The judge may detect this through a phrase that is almost natural. The words are grammatically correct. Their opposition is deliberate. The intended joke can be explained. Yet the sentence feels as though it was built backward from a desired punchline rather than spoken forward from a human situation.

That "almost" is enough. In an ordinary content task, a nearly natural sentence may count as success. In humor, the slight displacement becomes the entire result. Instead of laughing, the listener sees the generator.

7. Why N=1 Changes the Turing Test

Turing′ s imitation game is commonly discussed in terms of an interrogator′ s ability to distinguish machine from human through conversation. The numerical passage often associated with the test was Turing′ s prediction concerning an average interrogator under a limited interaction, not a universal law that defined one permanent pass mark.

The N=1 version removes the average. There is one judge.

This produces a fundamental change. The result can no longer be interpreted as a general percentage of people who might be deceived. It becomes a statement about whether this system can remain indistinguishable to this person.

If the sole judge is easily satisfied, the system may pass that encounter. If the sole judge has an unusual

linguistic ear, recognizes template structure, or shares no relevant humorous subset with the system, it may fail. The machine has not changed; the judge has. Therefore, the verdict does not belong to the machine alone.

In a strict N=1 test, the important requirement is not one successful answer. It is sustained indistinguishability. An isolated laugh proves only that one expression reached the judge. Humans laugh at mistakes, absurd failures, accidental word combinations, and the spectacle of a machine trying too hard. The fact of laughter does not by itself establish that the judge experienced the speaker as human. Conversely, the absence of laughter does not establish a machine: human humor also fails.

The judge can continue. Each new humorous attempt reveals more of the generator′ s habits. Repeated reversals expose a template. Excessive explanation exposes uncertainty. Sudden changes of style expose adaptation. Safe generality exposes the absence of a shared private center. The test becomes harder over time because the judge is learning what to look for.

Thus, a particular N=1 test can be passed or failed, but passage cannot automatically be universalized. To claim universal passage, one would have to show that the same system can sustain indistinguishability before any possible sole judge. Yet the decisive subset is defined by the unique judge, changes during interaction, and may contain private relations that no population model possesses in advance.

The very identity of the judge enters the criterion. Once this occurs, ″the AI passed the Turing test″ is no longer a complete sentence. Passed for whom, under what shared history, and for how long must be specified.

8. Why Humor Is a Decisive N=1 Probe, Not the Verdict

Many tests ask a system to display knowledge. Humor asks it to display knowledge of the listener.

A factual answer can be correct even when the system knows nothing about the person asking. A humorous answer cannot be fully correct in that way. It must choose not only content, but distance, permission, timing, compression, and voice.

The single judge can therefore probe dimensions that broad benchmarks average away:

- whether the system can create and preserve private meanings;
- whether it understands when a shared reference should remain implicit;
- whether it can distinguish natural speech from technically valid wording;
- whether it recognizes when its own previous failure has changed the atmosphere;
- whether it can abandon a learned comic form instead of merely varying its surface;
- and whether it can produce an insight that belongs to this conversation rather than to the general category of jokes.

This does not make humor mystical. It makes humor relationally exact.

The difficulty is not that a machine lacks access to a supernatural human substance. The difficulty is that the target exists in the narrow alignment between two changing systems. General training can bring the model close. Personal history can bring it closer. But the final criterion remains local: did these two spaces align on the decisive subset at this moment?

Under N=1, a local humorous failure cannot be averaged out by other judges who enjoyed the joke. But it becomes a Turing-test failure only if the sole judge can use it, usually as part of a repeated pattern, to distinguish the AI reliably. The verdict concerns discrimination, not laughter.

9. Conclusion

Humor is not the production of a joke-shaped sentence. It is the momentary, high-level, complete alignment of two minds over the small subset that decides how an unexpected expression will be understood.

That is why humor differs from ordinary functional approximation. A function can often be achieved through many routes and judged by a stable external result. Humor has no listener-independent correct output. Its target is selected by a particular person, in a particular relation, at a particular moment.

The same fact explains both the power and the limit of the N=1 Turing test. With one judge, passage cannot be established by average performance. The system must remain indistinguishable to the only mind whose judgment matters. One unnatural word, one visible template, or one failed implication may give that judge evidence, but no single bad joke is automatically a verdict.

An AI may make the judge laugh, even repeatedly, while remaining obviously artificial. In that case humor succeeds and the Turing test fails. An AI may also make a judge laugh while remaining indistinguishable; in that case it passes that judge's N=1 test. Conversely, a human may fail to make the judge laugh without ceasing to be human. The criterion of passage is therefore sustained indistinguishability, not laughter.

Because the decisive subset belongs to a unique relationship and changes as the relationship unfolds, passage before one judge cannot become a universal pass. Humor reveals this structural limit because it is an unusually sensitive probe of person-specific alignment:

> ***Humorous success and Turing passage are different judgments. The AI can pass for one judge and fail for another. A humorous mismatch counts against the AI only when it enables reliable identification. Therefore, a local N=1 passage cannot become a universal passing certificate.***

References

Clark, H. H. (1996). *Using Language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

Jentzsch, S., & Kersting, K. (2023). ChatGPT is fun, but it is not funny! Humor is still challenging large language models. *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 325-340. https://doi.org/10.18653/v1/2023.wassa-1.29

McGraw, A. P., & Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological Science, 21*(8), 1141-1149. https://doi.org/10.1177/0956797610376073

Rosenbusch, H., Evans, A. M., & Zeelenberg, M. (2022). The relative importance of joke and audience characteristics in eliciting amusement. *Psychological Science, 33*(9), 1386-1394. https://doi.org/10.1177/09567976221098595

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

任何机器都无法普适通过图灵测试

幽默作为个体化对齐的最终测试

严格单人裁判条件下高层子集对齐的概念解释

贾宝龙

独立研究者

双语概念论文·草稿 v0.4 ·2026 年 7 月

摘要

人工智能通常按照功能进行评价：它能否回答、预测、规划、模仿或生成。幽默却不同。一个表达并不会仅仅因为具备笑话的可识别结构，就成为成功的幽默。只有当说话者与某一个具体听者，在某一瞬间，就双方内部概念空间中一个狭窄而高层的子集完成对齐，幽默才真正发生。

本文把这一条件称为“**高层子集参数对齐**”。相关子集可能包含文化记忆、个人经验、未说出的预期、语言直觉、社会距离、情绪状态、禁忌边界、对话节奏，以及双方对当前关系的共同认识。两颗心智在整体上不必相同，但在决定这次幽默事件的子集内，对齐必须完整：一个错误暗示、一个不自然的词，或者一个错位的节拍，都可能摧毁全部效果。

本文区分三种绝不能混淆的判断。**幽默成功**是指一个表达使裁判发笑。**N=1 局部通过**是指经过持续互动以后，唯一裁判仍然不能可靠地区分 AI 与人。**普适通过**则要求同一个系统在任何可能的单一裁判面前都能实现局部通过。发笑既不是通过图灵测试的充分条件，也不是必要条件。裁判可以连续发笑，同时仍然知道对方是机器；此时幽默成功，但图灵测试失败。人类也可能讲出不好笑的笑话；此时幽默失败，却不能仅凭这一点判定说话者是机器。

幽默的重要性在于它可以成为探针。如果唯一裁判反复发现模板结构、不自然措辞或个体化对齐失败，并能够据此可靠识别 AI，系统就没有通过这个裁判的 N=1 测试。同一个系统仍可能通过另一个裁判的测试。因此，它可以局部通过，却没有普适通过。本文不是说机器永远不能输出好笑的句子，也不是说任何失败笑话都能证明机器身份，而是说幽默会暴露依赖裁判的对齐，使一次局部成功无法成为普适通行证。

关键词：幽默；N=1 图灵测试；人工智能；高层对齐；子集参数；人机互动；喜剧直觉

---

1. 错误的问题：机器能生成笑话吗？

理解机器幽默最容易犯的错误，是询问机器能不能生成笑话。

当然能。一个接受了足够语言训练的系统，可以复现幽默的可见形式：铺垫与包袱、反转、夸张、双关、不协调、自嘲和反讽。它也可以把这些形式组合成让一部分人觉得好笑的句子。但这些能力都还没有触及幽默最困难的部分。

笑话是语言对象，幽默是心智之间发生的事件。

对象可以被复制、储存、翻译和评分。事件却取决于谁在说、谁在听、双方共享什么、什么没有说出来、刚刚发生了什么、什么不能直接说，以及说话者在说出它之前停顿了多久。同一句话在不同关系中，可以是精彩、普通、残酷或毫无意义。

所以真正的问题不是：

> 系统能否构造出具有笑话形式的东西？

而是：

> 系统能否找到那个精确的隐藏区域，使这个具体听者把表达同时体验为意外而又正确？

这不是一般生成问题，而是针对具体个人的对齐问题。

2. 三种绝不能混淆的判断

讨论 AI 幽默时，“成功”实际上可能指三种不同结果。

幽默成功：听者产生愉悦或发笑。这描述的是一句表达造成的效果。

N=1 局部通过：一个指定裁判经过持续、开放式互动以后，仍然不能可靠地区分 AI 与人。这描述的是一次单人测试的结果。

普适通过：同一个系统面对任何可能的单一裁判，都能实现局部通过。这是跨越许多次 N=1 测试的主张，而不是某一次 N=1 相遇本身的结果。

可以用两个裁判把区别说清楚：

- 甲连续发笑，但知道说话者是 AI：AI 实现了幽默成功，却没有通过甲的 N=1 图灵测试。
- 甲发笑，并且在持续互动后仍不能可靠判断说话者是人还是 AI：AI 通过了甲的 N=1 测试。
- 乙从生硬或套路化幽默中识别出机器，并能够据此可靠判断：同一个 AI 没有通过乙的 N=1 测试。

综合结论非常明确：系统对甲通过、对乙失败。它既不能在不加限定的情况下被宣布为“通过了图灵测试”，也不能被说成每一次测试都失败。它通过了一次局部测试，又失败于另一次，所以没有普适通过。

幽默是探针，不是判决。一个差劲笑话本身什么也证明不了，因为人类也会讲差劲笑话。只有当幽默表现形成可重复的识别依据，使唯一裁判能够可靠地区分机器与人时，它才进入图灵测试的判决。

3. 功能近似与子集对齐

许多智能行为可以通过功能近似实现。系统接受任务与反馈，不断改善产生目标结果的能力。内部可以存在许多不同路径，只要功能实现，近似就算成功。

幽默的成功条件不同。不存在一种脱离听者的、名为“正确笑话”的输出。目标会随着接收者改变。

系统可以通过学习哪些表达经常使人发笑，来近似成功幽默的一般形式。这会产生人群层面的能力。它可以学会铺垫先制造一种预期，再由包袱用另一种解释取代它；也可以学会常见喜剧题材、安全的夸张幅度和熟悉节奏。因此，生成式幽默完全可以结构完整，也可以偶然有效。

但人群规律是对差异求平均，而针对一个人的幽默，恰恰取决于被平均掉的那些差异。

决定性问题不是系统是否一般性地学会了喜剧功能，而是系统与听者是否在此刻真正重要的那一小组参数上完成对齐。这个子集可以包括：

- 只有双方知道的一段记忆；
- 一个普通词在两人之间形成的私人含义；
- 从来没有明说过的预期；
- 这个人对大胆与冒犯的私人分界；
- 符合此人口语习惯的节奏；
- 某件新近事件所具有的情绪权重；
- 听者知道说话者正在试图制造幽默；
- 说话者也知道听者正在检验这次尝试。

功能近似允许替代方案；子集对齐却不允许关键位置上的替代。在当前活跃的幽默子集内，一个看似微小的偏差就可能决定全部结果。

4. “完全对齐”究竟是什么意思

“完全对齐”很容易被误解。它不是说说话者与听者必须拥有相同的全部心智。两个人可以在绝大多数事情上意见相反，却仍然共享一个完美笑点。

“完全”只作用于当前活跃子集。

假设一个幽默时刻依赖六个隐藏条件：双方必须认出同一段旧事、理解同一个双关、把同一规范视为暂时可以违反、感受到同一亲密程度、听出同一种反讽语气，并预期句子在同一个位置停止。如果五项对齐而第六项没有对齐，笑话仍可能完全失败。

所以，摧毁幽默的往往是极小的东西：

- 某个词在语义上说得通，但现实中的人不会在此处这样说；
- 一个反转可以被理解，却明显是强行制造的；

- 一个暗示几乎正确，却给人物安上了错误动机；
- 句子出现时，听者已经提前猜到结尾；
- 说话者解释了本应由听者自己发现的关系；
- 听者认出了违反，却没有把它理解为良性；
- 或者听者感到系统是在瞄准幽默模板，而不是响应当下情境。

失败程度与误差大小并不成比例。一个词就可以让整个事件坍塌，因为这个词恰好位于决定性子集内部。

这就是幽默作为高层对齐问题的含义：语法和字典意义都可以正确，表达整体却仍然错误。真正的判断发生在意图、关系、自然度、预期与共同暗示的层面。

5. 两个内部空间的瞬间交集

“向量空间”在这里是一种概念图像，不是数学证明。一个人的心智包含庞大的联想、记忆、价值、情绪权重、语言习惯与预测网络；另一个人拥有不同的网络。

普通交流要求部分重合。幽默要求一种更狭窄、更精确的状态：两个网络必须在同一瞬间，就正确的高层关系形成交集。

说话者首先把听者引向一种预期解释，随后通过幽默转折打开第二种解释。转折要想成功，听者必须能够从第一种解释移动到第二种解释，却又不能被完整告知中间路线。第二种解释过于明显，就没有闪光；距离太远，就只剩困惑；跨过错误边界，就变成冒犯；如果必须解释，时机已经失去。

因此，成功的幽默是一种带有张力的交集。说话者与听者必须对齐到足以共享隐藏路线，但听者又不能在转折出现以前就走完这条路线。

这也是为什么最好的幽默往往难以复述。它的信息内容可能很简单，真正的效果却存在于识别路径中。解释只给出目的地，无法重新制造移动过程。它能够说明笑话是怎么构造的，却无法复现两个内部空间相遇的瞬间。

6. “闪光”与模板幽默的失败

AI 幽默经常拥有结构，却没有闪光。

“闪光”不是随机、复杂或使用罕见词汇，而是一种特殊瞬间：一句话在出现以前出乎意料，在出现以后却显得无比准确。听者没有预测到它，却立刻感到它属于这个场景。

模板幽默把这种关系倒了过来。它的结构在意义真正活起来以前就已经暴露。听者先看见机器：正常铺垫、夸张反差、最后一句与开头形成对立。结果仍然可能可以理解，甚至有一点趣味，但机制比洞见更早到达。

大语言模型尤其擅长生成这种结构正确的幽默，因为它们能够学习语言形式之间反复出现的关系。同一项能力也制造了可识别的弱点。当许多候选模式被压缩为一个普遍安全的回答时，表达就容易像过去笑话的统计中心。它符合类别，却不符合这个人。

裁判可能通过一句几乎自然的话识别这一点。词语没有语法错误，反差也是故意的，笑点甚至可以被解释；但整句话让人觉得，它是从预设包袱反向拼装出来的，而不是从一个真实人物的处境中自然向前说出的。

这个“几乎”已经足够。在普通内容任务中，几乎自然可以算成功；在幽默中，轻微错位本身就会成为全部结果。听者不再发笑，而是看见生成器。

7. 为什么 N=1 改变了图灵测试

图灵的模仿游戏通常被理解为：询问者能否通过对话把机器与人区分开。后来经常与图灵测试联系在一起的数字，是图灵针对普通询问者和有限互动所作的预测，而不是一条定义永久及格线的普遍法则。

N=1 版本去掉了平均值，只保留一个裁判。

这造成了根本变化。结果不能再被解释为某个比例的人可能受骗，而只能表述为：这个系统能否在这个人面前保持

不可区分。

如果唯一裁判容易满足，系统可能通过这一次相遇；如果唯一裁判具有特殊语言直觉，能够识别模板结构，或者与系统没有相关的幽默子集，系统就可能失败。机器没有改变，裁判改变了。因此，结论并不只属于机器。

在严格 N=1 测试中，真正的要求不是一次回答成功，而是持续不可区分。一次偶然发笑，只能证明某个表达碰到了裁判。人也会因为错误、荒谬失败、偶然词语组合，或机器过分努力的样子而笑。发生了笑，并不自动证明裁判把说话者体验为人。反过来，没有发笑也不能证明对方是机器，因为人类幽默同样会失败。

裁判还可以继续测试。每一次新的幽默尝试都会暴露更多生成习惯。重复反转会暴露模板，过度解释会暴露不确定，突然换风格会暴露适应痕迹，安全而通用的表达会暴露双方之间缺乏私人中心。测试随着时间变难，因为裁判正在学习应该寻找什么。

所以，某一次具体的 N=1 测试可以通过或失败，但这个结果不能被自动普适化。要声称普适通过，就必须证明同一个系统在任何可能的单一裁判面前都能持续保持不可区分；然而，决定性子集由独一无二的裁判定义，会在互动中不断变化，还可能包含任何面向人群的模型事先不具备的私人关系。

裁判身份已经进入判定标准。到了这一步，“AI 通过了图灵测试”不再是完整句子。必须继续说明：通过了谁的测试、基于什么共同历史、持续了多久。

8. 为什么幽默是重要的 N=1 探针，而不是判决本身

许多测试要求系统展示知识；幽默要求系统展示它对听者的知识。

一个事实回答即使完全不了解提问者，也可以正确；幽默回答却不能以同样方式完整正确。它不只需要选择内容，还必须选择距离、许可、时机、压缩程度和说话声音。

因此，唯一裁判可以探测被大型基准平均掉的维度：

- 系统能否创造并保存私人含义；
- 是否理解共同典故什么时候必须保持隐含；
- 能否区分自然口语与技术上成立的措辞；
- 能否意识到自己上一次的失败已经改变气氛；
- 能否真正放弃一个学会的喜剧形式，而不只是改变表面词汇；
- 能否产生属于这段对话的洞见，而不是属于“笑话”这个一般类别的产品。

这并没有把幽默神秘化，而只是揭示了幽默在关系上的精确性。

困难不在于机器缺少某种超自然的人类物质，而在于目标存在于两个变化系统之间的狭窄对齐之中。一般训练可以使模型靠近，个人历史可以使它更靠近，但最终标准始终是局部的：这两个空间是否在这一瞬间，就决定性子集完成了对齐？

在 N=1 条件下，局部幽默失败不能被其他觉得好笑的裁判平均掉。但是，只有当唯一裁判能够利用这种失败--通常是利用反复出现的失败模式--可靠识别 AI 时，它才构成图灵测试失败。判决对象是能否区分，而不是是否发笑。

9. 结论

幽默不是制造一句具有笑话外形的话，而是两颗心智在决定意外表达将如何被理解的那个小型子集上，瞬间完成高层而完整的对齐。

这就是幽默与普通功能近似的不同。一个功能通常可以通过多种路径实现，并由稳定的外部结果评价。幽默却不存在脱离听者的正确输出。目标由一个具体的人、处于一种具体关系和一个具体时刻共同选定。

同一个事实也同时说明了 N=1 图灵测试的力量与极限。只有一个裁判时，平均表现无法建立通过。系统必须在唯一有判决权的心智面前持续保持不可区分。一个不自然的词、一个可见的模板或一个错误暗示，可能为裁判提供识别证据，但任何一个差劲笑话本身都不是自动判决。

AI 可以使裁判发笑，甚至连续使其发笑，同时仍然显得明显是机器。此时幽默成功，图灵测试失败。AI 也可能使裁判发笑，并且始终没有被可靠识别；此时它通过了这个裁判的 N=1 测试。反过来，一个人类可能没有使裁判发笑，却不会因此失去人类身份。因此，通过标准是持续不可区分，而不是发笑。

由于决定性子集属于一段独特关系，并随着关系展开而变化，所以通过一个裁判不能成为普适通过。幽默之所以揭示这一结构性极限，是因为它是个体化对齐的高度敏感探针：

> **幽默成功与图灵通过是两种不同判断。*AI* 可以对一个裁判通过、对另一个裁判失败。只有当幽默失配使裁判能够可靠识别 *AI* 时，它才构成测试失败。因此，一次局部 *N=1* 通过不能成为普适通过证书。**

参考文献

Clark, H. H. (1996). *Using Language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

Jentzsch, S., & Kersting, K. (2023). ChatGPT is fun, but it is not funny! Humor is still challenging large language models. *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 325-340. https://doi.org/10.18653/v1/2023.wassa-1.29

McGraw, A. P., & Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological Science, 21*(8), 1141-1149. https://doi.org/10.1177/0956797610376073

Rosenbusch, H., Evans, A. M., & Zeelenberg, M. (2022). The relative importance of joke and audience characteristics in eliciting amusement. *Psychological Science, 33*(9), 1386-1394. https://doi.org/10.1177/09567976221098595

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433